ADVISORY PANEL ON CONSUMER PRICES – STAKEHOLDER

Consumer Prices – Alternative Data Sources Roadmap

Status: (final)
Expected publication: (alongside minutes)

Purpose

1. The Alternative Data Sources Roadmap presented to APCP-S in January 2019 detailed our timeframe for implementation of new methods and data sources into headline consumer price statistics. This paper provides an update on progress made and our plans over the next 3 years.

Actions

2. Members of the Stakeholder Panel are invited to:
   a) provide feedback on the progress to date
   b) comment on the timelines and scope of the alternative data sources project

Background

3. We are currently working through a comprehensive transformation programme for consumer price statistics in order to modernise their measurement and make better use of data and methods that are becoming increasingly available to us.

4. At a high level, this involves obtaining robust sources of alternative data, development of statistical systems to work with these data, and methodological research in order to effectively classify, validate and construct high quality price indices from these new data sources. These new data sources will be used in conjunction with traditionally collected data to improve the accuracy, efficacy and representativity of consumer price inflation statistics.

5. The data sources we are investigating are web-scraped data (automated data collection from retailer websites) and scanner data (point-of-sale expenditure and quantity data provided directly by retailers). More information can be found regarding these data sources in our article Introducing alternative data sources into consumer price statistics.

6. This transformation will be the largest change to consumer price statistics in a generation, and the scale and importance of this work should not be underestimated. We will be reliant on developments in many areas, including the use of new technology platforms and the willingness of retailers to provide us point-of-sale data.

7. In this paper we provide a summary of our progress made over the last 12 months, and an update on our roadmap between now and 2023, when we plan to first incorporate data from alternative sources into our headline measures of consumer price statistics.

8. During this period, we will be liaising regularly with both the APCPs, our users, and the Office for Statistics Regulation, to ensure that our future plans for consumer price inflation measurement are appropriate for improving the quality of our statistics and meeting our ongoing user requirements.

Progress made in 2019

Obtaining robust sources of alternative data

9. We have been engaging with several retailers throughout 2019 and have made substantial progress in accessing their point-of-sale (POS) data. We now have test datasets from a number of the UKs biggest grocery retailers. We have initiated a regular data feed with one retailer and expect to begin receiving regular feeds for other retailers in the coming months. We have also gained access to historical datasets that are currently being used to complete research and impact analyses into use of the data.

10. Since November 2018 we have been receiving regular web scraped data from mySupermarket. These data are for around 25 categories in the basket covering areas such as clothing, electronic items and package holidays. This contract is now due to terminate in March 2020 and we are looking to re-procure imminently. The new procurement will have more of a focus, concentrating on items as prioritised with the APCP-S in September 2019. It should also enable us more flexibility to make modifications to the contract throughout the term to meet our ongoing requirements. There are no historical series available with these data so we will need to build up a sufficient time series of high-quality data before a final impact assessment can be completed.

11. While building up a time series of web-scraped data from external providers, we are also beginning to scale up our capabilities and systems to be able to web-scrape data in-house. We have built a web-scraping pipeline that will allow us to build and maintain in-house scrapers and are working to review the policies and systems in place to allow us to do this. This month we have started a collection of motorbike prices using web-scraped data in parallel to the traditional collection and plan to replace this in the live index production system for consumer prices in 2021.

Development of statistical systems

12. In 2019 a prototype pipeline was further developed to process web-scraped and scanner data on a strategic, distributed processing platform. The pipeline uses a modular approach to produce price indices; beginning with validation and classification of the data then averaging the data across the month, calculating an array of price indices using different elementary aggregate index number formulae and producing index outputs that can be analysed and compared between retailers and product types.

13. Scanner data were processed through the pipeline for the first time in 2019, showing the new systems capability to pass over 900 million rows of data through the pipeline to produce index outputs. We also published some experimental web scraped indices using this pipeline, showing the capability of the system to process several scenarios for research purposes, for example testing the impact of switching off outlier detection methods, or using different index methods.

14. The pipeline has also been developed further to include additional functionality in order to facilitate our research programme. This includes development of a filtering mechanism to select only products with a significant turnover, and the capability to calculate price relatives for groups of products, rather than individual products.

Methods research

15. Throughout 2019 we have continued to explore new methods and techniques for use with big data sources:
    - A framework has been developed to decide on appropriate index number methods for use with new data sources (as discussed with the APCP-T in January 2020).
    - Research into classification methods has resulting in us using a rules-based classification model for technological goods and the development of an application to enable quicker and more accurate manual labelling of data in order to train machine learning models for more items.
    - We have also investigated different ways that expenditure can be approximated for web-scraped data to ensure no products with little to no expenditure are having undue influence over the resulting price index.

Roadmap to 2023

16. We have split our delivery programme until 2023 into three phases, each lasting a year. In 2020 and 2021 we will discuss ongoing research and developments with both the Technical and Stakeholder Panels at regular intervals and will publish bi-annual research papers to update users on our progress. In 2022 we enter an engagement phase where we will be able to share aggregate quarterly experimental estimates incorporating alternative data sources.

17. The first phase (**research**) will run until the end of 2020 and involves:
    - continued engagement with retailers to secure more regular data feeds as well as historic data
    - integration of traditionally collected data into the processing pipeline and understanding the impact of processing scanner data, web-scraped data and traditional data simultaneously
    - Research into improvements that can be made to processing of traditional data including developments of new systems
    - further developments of the processing pipeline for web-scraped and scanner data to enable research and impact analysis
    - research into the methods needed to produce high quality indices using web-scraped and scanner data (a full research programme is outlined in **Annex A**)
    - a parallel run of using web-scraped data in a production environment
    - initiation of a new and more focussed web-scraping contract
    - a review into policy and system changes needed to fulfil the longer-term strategy to bring web-scraping in-house
    - user engagement through bi-annual publications and workshops

18. The second phase (**application**) will run throughout 2021 and involves:
    - application of research to specific item categories within the inflation basket as prioritised with APCP-S in September 2019
    - completion of approved methods built into the processing pipeline
    - initial impact assessments carried out on aggregate measures

- continued engagement with retailers to expand the amount of scanner data available to us and ensure continuation of regular data feeds
- scaling-up of in-house web-scraping capabilities
- a summary of research and final recommendations on methods for different commodity groups made
- planning priority items for beyond 2023
- user engagement through bi-annual publications and workshops

19. The third phase (**engagement**) will run through 2022 and involves:
    - quarterly publication of aggregate experimental indices including web-scraped and scanner data in conjunction with traditionally collected data
    - user engagement to discuss methods and changes
    - research and developments for priority items for beyond 2023
20. A visualisation of our high-level roadmap can be found in **Annex B**.

Prior to 2023 – some quick wins

21. There are several options for using web-scraped and scanner data in consumer price indices prior to 2023, but these will utilise existing methods and systems.
22. As already mentioned, there may be opportunity to include web-scraped data in production prior to 2023. This January we have initiated a parallel run of data collection for motorbikes using web-scraped data and are hoping to replace this in live production in 2021. We also may increase use of the "robot-tool" throughout 2020 and 2021, which monitors a website and sends a notification when changes to the website, or price, are made. The robot tool will be utilised where prices are relatively stable, such as for passport fees and birth, marriage and death certificate fees.
23. There is also a case for using web-scraped price and attribute data in the creation of hedonic models for technological goods. These models are currently based on data collected manually once per quarter; use of web-scraped data would ensure that models could be produced in a more timely and efficient way.
24. Scanner data may be used prior to 2023 to give us more prescriptive weights at the lowest levels of aggregation where there is little other information available (for example, expenditure on t-shirts relative to expenditure on shirts).
25. Finally, we may be able to apply take-up rates of multibuy discounts to traditional collection items prior to 2023, to ensure that we are closer to representing actual transaction prices.

Helen Sands
Prices Division, ONS
January 2020

List of Annexes

Annex A	Research programme for the alternative data sources project
Annex B	2020:2023 roadmap

Annex A – Research programme for the alternative data sources project

1. Research phase (2020)

Research into the methods needed to produce high quality indices using web-scraped and scanner data:

- **Classification** techniques for alternative data sources (including machine learning methods): This will enable us to automatically classify large quantities of data into price indices
- **Index number methods** framework: This will allow us to ensure we are choosing the most appropriate index number method dependent on the pricing behaviour of a particular item and characteristics of the dataset
- **Product grouping**: In areas with high product churn, product grouping methods will allow us to follow groups of products over time, rather than individual products
- **Scanner data research**: We need to better understand the scanner data before using it, e.g. how do we account for returns and discounts? Can we apply take-up rates of multibuy discounts to traditionally collected data? How do we identify product relaunches to ensure appropriate quality adjustment takes place?
- **Expenditure proxies**: This research will look at using proxies to weight web-scraped data to ensure that more popular products are given a higher weight in the index
- **Retailer weights:** We need to find data sources that will allow us to weight retailers together with traditionally collected data at the lowest levels of aggregation
- **Outlier detection and imputation:** This will enable us to appropriately identify and remove/validate outliers and identify appropriate imputation methods to handle missing data – whether the data is missing temporarily, permanently, or on a seasonal basis.

2. Application phase (2021)

Applying methods and techniques developed during phase 1 of the research programme to specific item categories. These item categories were agreed and prioritised with our stakeholders in September 2019.

- **Groceries**: With a focus on scanner data as we can cover a large market share with a small number of retailers.
- **Clothing:** With an initial focus on web-scraped data as market widely distributed across retailers. Clothing items have high product churn due to changing fashions and have led to serious problems with price indices in the past (e.g. formula effect).
- **Tech goods:** With an initial focus on web-scraped data as market widely distributed. Tech goods also have a high product churn and current hedonic methods are highly resource intensive.
- **Used cars:** Measurement in used car prices is challenging as need to adjust for age/mileage. Market engagement ongoing to procure a suitable data source that gives sufficient coverage of attributes to allow for quality adjustments.
- **Package holidays:** Current method of producing package holiday indices doesn't align with Eurostat methodology and is incoherent with calculations across the rest of the inflation basket. Initial focus on web-scraped data to provide a larger quantity and higher frequency of data collected.
- **Air fares:** Web-scraped data will provide more frequent collection over a significantly larger sample of routes.
- **Rail fares:** Current methodology uses an imputation based on the cap of regulated rail fares set by the chancellor. Scanner-type data will allow us to use actual transaction data to calculate the index
- **Chart collected items:** DVDs, CDs, Books, Computer games use a methodology that follows chart positions over time rather than individual products. As such the index can be volatile as products change position within the charts.

Annex B – 2020:2023 roadmap